A Report to the Library of Congress

WEB PRESERVATION PROJECT
FINAL REPORT

William Y. Arms
Cornell University

September 3, 2001

Contents

1. Background to the Web Preservation Project

This is the final report of work carried out on behalf of the Library of Congress during the period March 2000 to June 2001. As stated in the original proposal, the objective of the effort was, "... to initiate a broad program to collect and preserve open-access materials from the World Wide Web. The effort will include consensus building within the Library, joint planning with external bodies, studies of the technical and policy issues, the development of a long-term plan and coordination of prototypes."

For this purpose, the Library of Congress established a working team of Roger Adkins, Cassy Ammen and Allene Hayes from the Library Congress, with William Arms of Cornell University as consultant. Melissa Levine provided expertise on copyright and other legal issues. The team met regularly with an advisory group of Barbara Tillett, Jane Mandelbaum and Diane Kresh. This report makes extensive use of notes, comments and observations provided by all members of the team. However, the opinions expressed in this report are those of the author and not those of the Library of Congress.

In parallel with this work, the Library carried out an experiment with the Internet Archive to collect web sites associated with the Election 2000. The organizational aspects of this experiment are discussed in Section 2, below.

Overall, this work has successfully demonstrated the processes by which the Library can select, collect, organize and preserve open-access materials from the web by downloading copies over the Internet. There are no serious impediments to the Library undertaking a broad program of collecting web sites, as part of its mission to collect and preserve the cultural and intellectual artifacts of today for the benefit of future generations. However, this is a substantial undertaking, which would require a dedicated team of librarians and technical staff.

The Interim Report

This final report should be read in conjunction with the *Interim Report*, dated January 15, 2001. Material that is covered thoroughly in the *Interim Report* is not repeated in this report. It is available at:

- William Y. Arms, *Web Preservation Project: Interim Report*, January 15, 2001. http://www.cs.cornell.edu/wya/LC-web/interim.doc.

The *Interim Report* discusses the options for the Library of Congress and makes recommendations about how to organize a large-scale preservation program. In particular, the report describes the process by which web sites are collected and preserved, and the Minerva prototype to study these processes. It stresses that collecting, organizing, preserving and providing access to web materials are inter-related. Because of the volume of web materials that deserve to be collected and preserved, most processes will be automated, with skilled librarians establishing and monitoring the procedures.

As a result of the Minerva prototype, certain topics were identified as needing further study. They are discussed in this report. They include the role of partners and methods for collecting materials that are delivered over the web from repositories. In addition, the report proposes initial guidelines for developing policies for selection of web sites for preservation, copyright, access for researchers, and cataloging and indexing these materials.

Papers and other materials

Beyond the two reports, the Web Preservation Project created the following papers and online materials.

- Collecting and Preserving the Web: The Minerva Prototype, by William Y. Arms, Roger Adkins, Cassy Ammen, and Allene Hayes. *RLG DigiNews*, 5(2), April 2001. http://www.rlg.org/preserv/diginews/diginews5-2.html#feature1.

- *Minerva Web Site*. http://www.loc.gov/minerva. (Access restricted to within the Library of Congress.)

- *Minerva, the Web Preservation Project*, presentation by William Arms, Cassy Ammen, Allene Hayes, Jane Mandelbaum and Barbara Tillett to the Library of Congress, February 2, 2001. http://www.cs.cornell.edu/wya/LC-web/minerva.ppt.

The Election 2000 Collection is at:

- *Alexa Wayback Machine: Election 2000, an Internet Library*. http://archive0.alexa.com/collections/e2k.html.

2. Partnerships between the Library of Congress and Other Organizations

The Library of Congress has special responsibility to collect and preserve the nation's intellectual output for future generations, but it is not alone. As discussed in the *Interim Report*, the Library of Congress is likely to develop its preservation programs in conjunctions with many libraries, archives, publishers and other organizations. Recent Congressional funding has been explicit about the need for collaboration and has identified several likely partners, while requesting the Library of Congress to take the lead in overall planning. Here are some likely partners.

National libraries of other countries. These libraries emphasize their own national materials, but the web does not have boundaries. Among national libraries, the pioneers of collecting and archiving web materials have been the national libraries of Sweden and Australia. More recently, the National Library of the Netherlands has begun to develop a vigorous program. The British Library has been slow to develop a strategy for the web, but is now in a period of rapid expansion.

Federal agencies. Within the United States, several other federal agencies have missions that complement the Library of Congress. They include the National Library of Medicine and the National Agricultural Library, and the National Archives and Records Administration.

Research libraries. As yet, research libraries and archives in the United States have paid little attention to collecting web materials. For example, few if any universities systematically archive their own web sites. However, almost every major library has some collections in digital forms that it feels an obligation to preserve, often converted from physical materials. Academic libraries have a tradition of collaboration and, despite the slow beginnings, are likely to be dependable long-term partners.

Publishers. In the past, publishers have not seen preservation as one of their functions, but there are signs that this is changing. The agreement between the Library of Congress and UMI (now Bell and Howell) was the first in which the Library explicitly designated a publisher as the definitive archive for important materials. More recently, the American Physical Society has been collaborating in developing a shared preservation plan.

New organizations. New organizations are emerging that have preservation as a major function, which they perform for the common good or for the scholarly community. Two interesting but completely different organizations are JSTOR and the Internet Archive.

Coordinating and managing such partnerships is a significant task. The Library will need its own skilled staff to oversee the overall goals and to monitor the effectiveness of the collective efforts.

Preservation agreements

Many organizations and individuals will collect and preserve materials independently of any relationship to the Library of Congress. Indeed this independence is important for preservation. If different organizations, in different countries with different cultures, carry out different preservation programs the materials that they collect are vulnerable in different ways. This diminishes the risk of everything being lost through a single disaster or mistakes of technology and organization.

However, there are even greater benefits from coordinated planning and the world looks to the Library of Congress to take the lead in establishing and maintaining partnerships. For preservation partnerships to be of value to the Library of Congress, the Library must consider three criteria.

<u>Dependability</u>. Within a defined scope, the Library must be confident that the partner will carry out its agreed tasks. The scope will inevitably have restrictions and will have different dimensions for different partners.

<u>Resources.</u> The partnership must be beneficial to the Library. Frequently the benefit will be cost savings, particularly when a partner carries out preservation tasks with its own resources. On other occasions, the benefit will be expertise in content or technology. With the vast variety of forms and formats of digital information being continually introduced, the Library will often wish to turn to specialists rather than develop its own expertise in every category.

<u>Content.</u> Some categories of material deserve to be preserved but are not available to the Library. An important example is trade secret information, such as the source code of computer programs. It may be possible to reach agreement with the owners of such materials that they will preserve them, even if they will not trust any other organization to hold them.

These criteria will require trade-offs. For instance, agreements with commercial partners are subject to all the vagaries of commercial organizations that are at the mercy of the short-term outlook of the stock market.

As discussed in Section 5, under U.S. copyright law, the Library of Congress has special legal privileges to acquire materials from publishers and preserve them for the future. As presently written, this legislation does not explicitly permit the Library to delegate these functions to other organizations, but there are current discussions about asking Congress to allow such delegation. This would enable the Library to make use of special expertise or alternative sources of effort.

Some partnerships will be formal agreements with the Library of Congress. For instance, the Library has signed agreements with Bell and Howell, and with the Internet Archive. In particular, any situation where a third party acts as the agent of the Library under copyright law will certainly need to have a written agreement. Formal agreements on preservation partnerships need to cover certain topics:

Materials covered. Most partners will take responsibility for collecting and preserving a category of materials, e.g., university theses, open-access web pages in the *.gov* domain, publications of the American Physical Society, and so on.

Method of collection. How will the materials be collected? If the partner is not the owner of the materials, what permissions are necessary?

Relationship with the Library of Congress. The partner may be acting as the agent of the Library of Congress.

Copyright deposit. The CORDS system has the potential to permit various modifications of the traditionally combined tasks of copyright registration and deposit. The Library may agree for a publisher to preserve materials on behalf of the Library as an alternative to physical deposit.

Access. Partners must not only preserve materials. They must also provide access to Congress and to scholars. While the guidelines for access to digital materials are far from clear, if preservation partners are acting on behalf of the Library they should follow the same guidelines, or, where differences are necessary, the differences should be simple and comprehensible.

Risks. From the beginning of any partnership, precautions are needed to anticipate what can go wrong. Partners can go suddenly bankrupt or fail to carry out their tasks effectively. The Library needs to be confident that, if problems arise, it is able to take control of the preserved materials. Possible mechanisms include regular mirroring of the materials at the Library of Congress or deposit with an escrow agent.

All of this requires active involvement of Library of Congress staff. Partnerships cannot be created and left unattended. They require constant supervision and regular auditing.

The Internet Archive and the Election 2000 Collection

The collaboration between the Internet Archive and the Library of Congress is an excellent example of how partners with differing skills and resources can complement each other.

As a Silicon Valley insider, the Internet Archive has access to advanced technology and expertise. Web crawling, automatic indexing of web sites, and the techniques of managing massive collections on commodity hardware are well established among the best Internet companies, but have not percolated into the mainstream computer industry. Consequently, the Internet Archive's costs for collecting and preserving web materials are much lower than the Library's would be.

Because the Internet Archive has a simple decision-making structure – one person makes all the decisions – it has the flexibility to move quickly. It began collecting on its own initiative and its collections from the past five years contain vast amounts of web incunabula that would otherwise have been lost. The Library's decision making is constrained by long budget cycles, a less flexible structure, and close scrutiny by Congress and the public. However, the converse of flexibility is instability. The long-term future of the Internet Archive cannot be guaranteed. The organization itself may not have a long life and its focus on preserving the web may change. Its procedures for replication and backup of data are less stringent than, for example, Bell and Howell's archive of its digital collections.

Finally, in areas of legal uncertainty, the Library of Congress is very constrained and must avoid controversial activities. The Internet Archive is less constrained. It has chosen to take the steps that it believes are in the public's best interest, while recognizing that the legal framework is subject to varying interpretations. In general, the scholarly community gains greatly from the Internet Archive's willingness to accept legal risks and press ahead, but there is always a possibility that it could go too far and cause political or legal troubles for itself or its partners.

The Election 2000 Collection demonstrated many of the benefits of collaboration. The work involved two technical components that would not normally be available to the Library. The actual collection of materials used the Mercator web crawler developed by Compaq SRC. (This is the research part of the group that created Altavista.) The user interface provides access to the snapshots via Alexa’s *Wayback Machine* technology.

Using these tools, snapshots were collected from over 800 web sites between August 1, 2000 and January 21, 2001. Rapidly changing sites were archived daily or even several times in a day. They have been made accessible on the Internet, with a user interface that provides searches by date, by website, and by category. The total size is about two terabytes.

This experiment also showed some of the challenges of working with an independent partner. The level of staff time required was not realized initially; as the project developed the need to have a member of the Library's staff monitor the project became apparent. The agreement between the Library and the Internet Archive is informally worded, which led to some uncertainties, yet it would be sad if every small partnership needed a complex legal document before it began.

3. Collecting Materials from Repositories

Many web sites store collections of materials, such as documents, images and sound recordings, in back-end repositories. The usual architecture is to combine a front end that is implemented as a conventional web site with a repository of objects that are accessible through the web interface. The repository may be simply a set of Unix files, records in a relational database, or a specially designed system. The links between the front and back ends often use CGI scripts, where parameters to the scripts translate into the keys used to identify individual items in the repositories. This architecture is widely used with both open-access and restricted-access collections. For instance, some publishers combine an open-access front end with restricted-access to items in the repositories.

American Memory is an example. The web site of HTML pages and associated images provides the user interface, connections to the search engine, and screens for displaying results of searches and the content of the collections. The collections themselves and the searchable indexes are stored separately.

This is a good architecture for managing large collections. However, it poses difficulties for preservation, even when all the materials are available with open access. A web crawler or mirroring program will find only those pages that are referenced by explicit URLs, in practice the front-end web pages but not items in the back end. Yet, often, the most important content is in the back-end repositories.

A simple technique for collecting from repositories

Back-end repositories can be extremely complex or obscure in how they are constructed. For example, many scientific or medical databases cannot be interpreted without knowledge of the underlying schema, data dictionaries, and so on. Fortunately, many repositories have a simple structure. They consist of a set of records, each of which is a complete object such as a document or an image that can be retrieved from the web by using a well-defined URL. Thus, for example, every item in the American Memory collections can be retrieved by a specific URL.

Under these circumstances, if an archive or library has a list of all the URLs, it can collect and preserve the content without knowing how the repository is structured or how it is implemented. With this list, it is possible to collect the materials entirely automatically, using the URLs to download the items individually.

For access, the archive cannot attempt to replicate the internal structure of the original system. Instead, it should build its own version of the web site that was archived, by storing the items in any convenient data structure, indexed by the original identifiers. For access to these items, it is necessary to modify URLs in the preserved copy of the front-end web site, so that instead of linking to the original repository (perhaps through a CGI script) they connect to the new data structure. Depending on the construction of the URLs this can be anything from a simple, automatic process, to extremely complex. Fortunately, most big publishers manage their URLs very consistently, so that for many web sites this is a simple mapping that is kept constant over long periods of time. Often the syntax of the URL has two parts. The first part is a basic string that applies to all items. The second part provides an identifier for each individual item.

Thus the archive can preserve the individual items and provide limited access to them. The limit on access comes from the fact that the archive's data structure is not the same as the original back end. Usually, it will be straightforward to provide access to individual items, when used in a static manner, but in general it will not be possible to provide methods of access that, in the original web site, would have required manipulation, e.g., by a back-end database. For many collections this is not a significant restriction.

Cooperation with publishers

It is rarely possible to collect materials that are stored in repositories without having a list of the URLs needed to retrieve them. Unless such a list is embedded in a web page, this requires cooperation from the publisher. Therefore, it must be convenient for the publishers to provide lists of the URLs that access every item in the collections, preferably in the form of a basic string that applies to all items and a list of individual identifiers.

All parties would benefit if a standard mechanism were used to provide these lists. Developing this standard is beyond the scope of the current study, but there is one obvious candidate. The metadata harvesting protocol developed by the Open Archives Initiative provides a simple mechanism by which the managers of a collection can make metadata about its content available to others to harvest at a convenient time. To date, this protocol has been used primarily to exchange descriptive metadata for information discovery, but the protocol could be used for any category of metadata, including the lists of identifiers needed for preservation.

This method of collection can also be used to collect materials that are not available with open access, but require authorization for access.

4. Selection of Web Sites for Preservation

The basic method of collecting web sites is to copy the files from a web site to a computer at the Library of Congress, or some other library or archive. This is called a *snapshot*. The task of selection is to decide which materials to collect, the frequency of making snapshots, and related details of storing and indexing them.

At the beginning of the Minerva study, two principal methods of selection were proposed: *selective*, where librarians select individual sites based on their knowledge of the contents and expectations of future utility, and *bulk* where automatic web crawlers collect all materials that fall within very broad categories. The study has confirmed these two categories, but has also shown that there are many gradations between them that combine aspects of automatic collection with selection by professional librarians. For example, a librarian might define a category of material by some set of criteria but rely on automatic processes to identify the web sites that satisfy those criteria and collect them.

Minerva and the Election 2000 study also highlighted the fact that selection is much more than a simple decision whether to collect certain web sites. Decisions have to be made about the process of collection, the indexing, the organization of the collections for access, and the strategy for preservation. These decisions have large cost implications and have profound implications for the types of use that can be made of the materials.

There is no single correct decision. Low-cost bulk collection, as practiced by the Internet Archive, results in very large collections being preserved, but they are poorly organized, lack some categories of material and require skilled effort to use for research. They can be compared to boxes of unsorted papers deposited in an archive. Materials that have been carefully collected and organized, such as the Australian Pandora project or the Election 2000 Collection, are more complete and can be used immediately by scholars without special expertise. However, because they are labor-intensive to create and manage, their coverage is inevitably limited.

The converse of preservation is loss. What are the risks associated with these two strategies? With selective collection, anything that is not explicitly selected will be lost. This is likely to include materials from the early stages of a new organization, the youthful contributions of an individual who later becomes prominent, much of popular culture, and so on. The fact that all of these will be swept up by a comprehensive web crawler is perhaps the key reason why bulk collection should be a component of the national plan.

However, automation is no substitute for the judgment of selection librarians. Librarians can focus attention on especially important materials, or devise appropriate treatment for unusual materials. When a knowledgeable librarian selects web sites, decisions can be made about the following.

Frequency. Determining when to take snapshots is best done by librarians who have knowledge of the subject area. For example, snapshots of a quarterly magazine might be taken quarterly. For the Election 2000 project, snapshots were taken daily or even more often. Although some research has been done to identify the rate of change automatically, bulk collection usually chooses a standard frequency – perhaps monthly – for all sites.

Formats. The Internet Archive has been able to collect every open-access web site on a regular cycle by concentrating on a few formats, initially HTML pages, now augmented with related images. Other formats, such as audio, video, page images (e.g., PDF), and executable code, add greatly to the volume of material to be collected. Sometimes such materials are peripheral to a web site, but on other occasions they may contain vital content.

Repositories. As described in Section 3, many web sites depend on materials in repositories that cannot be collected without cooperation from the publisher or manager of the site.

The *Interim Report* discussed the interrelationship between selection, providing access to scholars, and long-term preservation. Decisions to collect all formats or to include material from repositories broaden the content available to scholars, but at considerable expense. Whereas the basic formats (notably HTML, JPEG and GIF) are very widely supported, other formats may depend on specialized or proprietary software.

The simplest form of access is to provide scholars with copies of the snapshot files, exactly as downloaded from the web. For some categories of research, access to this primary material is exactly what the scholar requires, but frequently the scholar would prefer to have an access version of the web site that is as close as possible to the experience of using the original web site at the time that it was collected. To create such an access version is not trivial. To maintain it over time in the face of technological obsolescence can be difficult or impossible.

Overlap with copyright deposit

This study has concentrated on open-access materials, but the boundary between open and closed access is fuzzy. The Library's special legal position can be applied to several areas where important materials are being lost or are in danger of being lost.

Online newspapers. Online newspapers are among the most important categories of information on the web. Although many are openly accessible by users, they cannot be collected automatically, because of robot exclusion, authentication requirements, or the use of back-end repositories.

Definition of best edition. Where materials are published in both online and paper forms, the two versions often diverge. Newspapers are a typical example. Currently, the Library collects the physical edition. The Library needs to be in a position to define as "best edition" the version that is genuinely best, or to collect both if they are very different.

Back-end repositories. Section 3 discussed back-end repositories. Much of the material in these repositories falls under copyright deposit and lies within the scope of the Library's collections.

Guidelines for selection policies

Here are initial guidelines that could be the basis for a collection development strategy.

Bulk collection. At present, it appears likely that the Library of Congress will decide not to carry out bulk collection of the entire web itself, yet bulk collection is one key component of preserving the web. Therefore, it is important to work with partners who will use automated methods to collect as much as possible of the web on a routine basis, preserve it and make it available to scholars as unprocessed files. One obvious partner is the Internet Archive, but there are other possibilities. For example, Google's monthly cache is currently discarded – at least partly for reasons of copyright – but executives of Google have privately expressed a willingness to consider presenting it to an archive. The Library can facilitate such efforts in several ways, the most important of which is to clarify the legal framework.

Selection of open-access web sites to be collected by the Library of Congress. Recommending officers across the Library know of many web sites that contain important materials that fall within the general scope of the Library's collections and are comparable in importance with the physical materials collected at present. They include online serials, political web sites, special events, and so on. Some materials are published only online; with others, the online version is the definitive version.

One simple principle applies to the selection of such materials. The Library should select materials based on content and the importance to the mission of the Library. The criterion should not be whether materials are in digital formats, but whether collecting and preserving them strengthen the Library's collections for future generations.

Use of legal deposit. When recommending officers select materials that cannot be accessed by downloading without the participation of the publishers, the Library should use its rights to acquire materials through legal deposit.

Selection of sites to be collected by partners of the Library of Congress. As discussed in Section 2, it is appropriate for the Library to enter into partnerships with libraries, archives, publishers and other organizations that are collecting and preserving web sites, or have a special interest or special responsibility for specific categories of material. For instance, the National Archives and Records Administration has responsibility for U.S. government records. In general, such partners can be expected to collect a restricted body of material, but to manage its preservation and access to a higher standard than is possible with bulk methods of collection and preservation. In aggregate these partners are likely to preserve more sites than the Library is able to collect.

Support for selection and preservation of sites by independent organizations. Many organizations that preserve web sites will not be formal partners of the Library. Indeed they may be unknown to the Library. The Library can help such organizations by establishing a framework of good practices for collection, preservation and access, and by creating guidelines for the use made of such materials.

5. Copyright

The legal issues in collecting open-access web sites were discussed in the *Interim Report*. As stated there, "For the Library to carry out its responsibility to preserve digital information – most of which is subject to copyright – the legal framework must be clear and unambiguous. While it is reasonable to assume that most organizations that make information openly available on the web would be willing for the Library of Congress to download copies and keep them for future research, the Library does not currently have the explicit legal right to do so."

The Copyright Office has offered to work with the Library to make explicit exactly what is required and, if necessary, to ask Congress to amend the Copyright Act to permit downloading of open-access materials that are on the Internet. This is important because some lawyers would argue that the archival activities of organizations such as the National Library of Sweden and the Internet Archive might go beyond the letter of the current law.

Guidelines for copyright policies

The *Interim Report* identified three areas where the Library needs new authority to implement a program of collecting and preserving open-access web sites:

<u>Downloading as an alternative to deposit.</u> Where materials have been made openly available without restrictions, the Library of Congress will download copies from the web rather than demand copies from the publisher. Moreover, in these cases, the Library will not ask permission before downloading materials for preservation.

<u>Use of partners.</u> The Library of Congress may choose to designate one or more other organizations, at locations other than the Library, to act as its agents to carry out collection and preservation of open-access materials on its behalf.

<u>Editing materials for preservation and access.</u> The Library will often make small editorial changes to the materials that it downloads for reasons of access and preservation. For instance, the Library might change an absolute URL to a relative URL, or a dynamic date to the date on which the item was collected.

In Section 4, above, there was a brief discussion of the importance of the Library extending its view of copyright deposit to include web materials. While all the suggestions in that section appear to fall within the Library's mandate under the current law, new regulations will be needed to implement them, such as revising the definition of "best edition."

6. Access for researchers

While the *Interim Report* made good progress towards clarifying the legal position in collecting open-access materials from the web, it did not address the question of who has access to the materials once they have been collected. The difficulty is that there are no good parallels to use in setting such policies.

An initial view might be that, since these materials were made available on the web with open access, the copyright owners expect them to be read and studied without restriction. For many sites, probably most, this is a valid assumption. The people who mounted them hoped that they would be read and will be pleased that they are available for future scholars. Other copyright owners, however, may be less enthusiastic. Materials may have been mounted for a specific purpose at a specific time; they may have included errors that were subsequently corrected. The potential for violations of privacy by data mining also needs to be taken seriously. While the aim should be to encourage widespread use of the collections, totally unmonitored access to all collections is probably inappropriate.

Conversely, a sad mistake would be to fall back on the habits of the past and require users to be physically present at the Library of Congress. This imposes burden on scholars by forcing them to travel unnecessarily and places burden on the Library to provide computing facilities for researchers to analyze the materials.

Guidelines for access by researchers

Here are some proposed guidelines for access by researchers.

<u>Registration of users.</u> Users of the collections would be registered. Access would be controlled by requiring users to login. The registration process could be quite simple, but users would need to sign a statement that they are using the materials for purposes within the Library's guidelines on use. To implement this policy, the Library would need to develop an online system for authentication of registered users.

<u>Guidelines on use.</u> The collections are available for research or education. They must not be used in ways that violate the privacy of individuals or the economic interests of the copyright owners. While commercial research is not excluded, the collections should not be used for every day operational purposes.

<u>Location.</u> Registered users may use the collections from any place on the Internet. Web sites are designed to be used by computers over networks. Scholars should be able to use the collections where it is most convenient for them to do research. Many researchers will wish to write computer programs to analyze the collections.

<u>Equipment.</u> The Library will provide a standard interface to the collections. Users will provide their own computers to connect to this interface. At a minimum, the interface

will allow the materials to be examined using a web browser, but the Library may also provide interfaces for analysis of the collections by computer program.

In devising policies and procedures for access, it is important to study the preferences and habits of users of digital libraries, not to model policies on those that have been developed for traditional collections.

7. Catalogs and indexes

Since very large numbers of web sites will be collected and preserved, some form of catalog, index, or finding aid is required, so that researchers can know what materials are held in the collections. Because there is no body of experience about how researchers will use a large archive of web materials, it is impossible to be explicit about the best form of indexing. Here are some general observations.

The Minerva study identified several characteristics of web sites that are troublesome for standards methods of indexing. Snapshots of a web site taken at different times can differ enormously. For instance, the sites of the presidential candidates changed dramatically when the vice-presidential candidates were chosen. Identification of web sites is a continual problem. URLs change frequently. A given web site may be referenced by several URLs, any one of which may change at any time. Conventional indexing categories, such as author and title, are often poor descriptors. For example, the title of a site may be poorly defined or not very useful as a search field. The HTML <title> can sometimes be used, but it is often inappropriate. Titles on the rendered version of the home page may change erratically.

Indexes to the web

There are two very different approaches to information retrieval of web materials: automatic indexing of the full text and manual cataloguing of web sites or items within sites. Although it is a new direction for the Library of Congress to take, the evidence suggests that automatic indexing of full text is likely to be the better choice.

Indexes, such as Google, combine a high-performance web crawler with a full text indexing and retrieval system. To achieve high capacity, they use parallel processing and run on banks of commodity hardware. The web crawlers have politeness algorithms to avoid overloading sites that are being indexed and to respect sites that do not want to be indexed. The retrieval methods all use some version of vector similarity. The differences are in their approaches to ranking and elimination of duplicates. These automatic systems have proved extremely popular with users, both scholars and casual users. Automatic indexing is, of its very nature, much cheaper than any form of catalog or index created by skilled professionals.

Efforts to extend conventional library catalogs to the web have had much less impact. OCLC's CORC project is an attempt to establish shared cataloging of web sites. For the Minerva study, MARC item level catalog records were created for each of the web sites, using the CORC software. The study showed that there are no fundamental obstacles to integrating such records into the ILS and the Library's other procedures, beyond the labor involved. The CORC project has created good software tools and a substantial number of web sites have been cataloged, but overall the project has not been adopted by many users. The Dublin Core metadata initiative has also had little impact on information discovery on the web, because very few web sites have consistent metadata. Yahoo! is the only widely used index that includes manually created records and it also has a large automatic index.

There are good reasons that users prefer the automatic indexes. First, as has been shown in a number of information retrieval experiments, when the full text is available, indexing of every term usually proves more effective for retrieval than searching surrogates such as catalog or indexes records. Second, because of its very low cost, automatic indexing can be much more comprehensive in coverage. Every word on every page of every item can be indexed. Low cost also means that, as materials change, the indexing can be repeated at regular intervals. Finally, untrained users find automated indexes more intuitive than conventional indexing services and catalogs, because they use natural language rather than controlled vocabulary, and ranked retrieval rather than fielded Boolean searching. In combination, these reasons all confirm that automatic indexing of full text is the most effective way to index web materials.

To carry out automatic indexing of its web collections, the Library would run an indexing program on the text files that have been collected. This creates a full text index of those files that can be searched directly by users. The entire collection can be indexed afresh at regular intervals, perhaps weekly. Several indexing programs are available commercially. For instance, the InQuery indexing program, which is used by American Memory, was used by Infoseek an early web search service. In this manner, the Library of Congress can develop its web preservation program without the need to build a large team of professional cataloguers.

Provenance metadata and indexes

Provenance metadata is essential for the preservation of web sites. This comes at several levels. Each file that is downloaded needs to have metadata that provides, as a minimum, the file name, the URL and IP address that it was collected from, the data type, and the date and time that it was collected. Since many versions of the file will be made for preservation and access, a unique identifier needs to be given to the file.

Each subsequent version of the file needs to have metadata that links it to the original and specifies what transformations of the file have been made. (The representation of transformations is one of the subjects of the OAIS model of preservation. Although the assumptions behind the model are different from the needs of web preservation, many of the OAIS concepts can be used.)

Over time, as repeated snapshots are made, a single file on the web may be collected many times. It is valuable to recognize that these have the same content. The standard method to do this is to calculate a unique code for each file, e.g., the MD5 hash, which generates a distinctive 128-bit code for each file. If two files differ by a single bit, they have different codes. If several files have the same code the probability that they are not identical is essentially zero. Only one copy needs to be preserved and only one copy needs to be indexed. The list of codes is a list of all the distinct files that have been collected. A searchable index of all the text in these files would provide access to the content of all the text files that have been collected. This provides an entirely automated way to index text files.

Automatically generated indexes will allow users to search the provenance metadata. These indexes will allow queries such as, "What versions of the web site *www.loc.gov* have been preserved from the year 2001?"

Guidelines for catalogs and indexes

Here are some recommendations for cataloging, indexing and other metadata for the collection of web sites.

Automatic indexing. The Library should rely on automatic indexing as the primary means for information discovery of web sites and content within web sites. The full text of all textual materials should be indexed on a periodic basis and searchable by users.

Cataloguing. The Library should not invest in extending MARC cataloguing or other manual methods of cataloguing to web sites, except for sites of particular importance to the Library and its users.

Non-textual materials. Files in formats other than text pose a problem. One optimistic approach is to hope that every non-text file belongs to a collection that is referenced by a text file and to reply on the textual indexes. This is probably not adequate for some of the more important sites. For these sites, it may be necessary to create collection-level catalog records, perhaps based on Dublin Core, or finding aids.

Provenance metadata. The generation and management of provenance metadata is an integral part of collecting and preserving web sites. The Library needs to work with other libraries and archives to develop standards for creating and distributing such metadata.